Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02427-x OPEN

Probability distribution of dependency distance and dependency type in translational language

Lu Fan [1,2]✉ & Yue Jiang[1]

As a "third code", translational language attracts considerable attention in linguistics research due to its distinctive features. Adopting the quantitative linguistic approach, the current study examines its features by investigating the mean dependency distance (MDD), as well as the probability distribution of the individual dependency distances (DDs) and distribution of a high-frequency dependency type in translational language. The MDD and the distributions were tested in a self-built corpus which contains parallel and comparable language materials in both Chinese-English and English-Chinese translations. The results show that: (1) compared with source texts and native texts, translated texts in both translation directions yield an MDD in between; (2) both the distribution of DDs and that of the dependency type *nsubj* follow the Zipf-Alekseev distribution in translated texts, as in source texts and native texts; (3) the in-between feature is further confirmed by parameters *a* and *b* in Chinese-English translation materials when fitting the distribution of DDs to Zipf-Alekseev distribution; (4) translational texts in both directions show higher *a* and lower *b* than their source and native texts when fitting the DD Distribution of dependency type *nsubj* to Zipf-Alekseev distribution. These findings suggest that, on the one hand, dependency distance minimization (DDM) occurs in translational language, which is consistent with native language and reflects a general tendency of natural languages to reduce cognitive load; on the other hand, translational language presents distinctive feature in *nsubj* type, but in most cases, it is subject to the gravitational pull of both source and target language systems, exhibiting a "compromise" feature in between. The current study highlights the contribution of syntactic quantitative methods to deeper understanding of the complexity of translational language and its cognitive underpinnings.

[1] School of Foreign Studies, Xi'an Jiaotong University, Xi'an, Shaanxi, China. [2] Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an, Shaanxi, China. ✉email: fanlu@xjtu.edu.cn

Introduction

As translational language arises from the process of rendering coded elements into other codes, it is referred as "third code" (Frawley 1984), "third language" (Duff 1981), "hybrid language" (Trosborg 1997) and "constrained language" (Kruger and Rooy 2016). It has aroused considerable attention from linguistic researchers, especially from those in descriptive translation studies. The distinctive features of translational language, also called "translation universals" (Baker 1996), have been depicted by researchers using a variety of indices, such as type-token ratio, sentence length, entropy and specific culture-loaded words or patterns (Xiao 2010; Laviosa and Liu 2021; Liu et al. 2022; Wang et al. 2023). Previous studies, however, have rarely examined the features of translational language through syntactic dependency analysis. Quantitative analysis of syntactic dependency structures can provide the study with holistic syntactic indicators which help to explain the cognitive factors underlying the distinctive features of translational language.

Syntactic dependency analysis is a method of analyzing syntactic features of language based on dependency grammar (DG). Dependency grammar represents unequal syntactic relationships between words and regards words attaching to each other in the way that one word is the governor (or head) and the other the dependent (Tesnière 1959; Liu 2009a; Hudson 2010). Dependency distance (DD), an indicator describing the linear distance between the governor and the dependent (Heringer et al. 1980), can be measured by counting the intervening words (Hudson 1995) or their linear position difference in a sentence (Liu et al. 2017). In the past decade, DD has become a popular indicator in quantitative linguistics due to its clear definition and underlying cognitive explanations. Quantitative studies on DD have been conducted on plenty of human languages, produced by either native speakers or second language learners (Ferrer-i-Cancho 2015; Ouyang and Jiang 2017; Wang and Liu 2017).

Recently, there have been two DG-related quantitative studies on translational language (Fan and Jiang 2019, 2020), in which the researchers intended to further verify the existence of distinctive features of translational language based upon dependency treebanks and dependency syntactic networks. In one of the studies, Fan and Jiang (2019) examined two syntactic indicators of translational language, i.e., mean dependency distance (MDD) and dependency direction, and compared the results with those of native language. The study found that there are differences between translational language and native language in terms of MDD and dependency direction. It is the first of its kind to introduce syntactic dependency analysis into the study of translational language and contributes to promoting translation studies through adopting quantitative methods and shed light on the cognitive process involved in translation activities.

However, there are still some gaps. One of them is the neglect of individual dependencies. Fan and Jiang (2019) mainly focused on the mean dependency distance rather than every single dependency in the treebanks. Since individual dependency distances provide more details of the fluctuation than the average which would level up differences of dependencies in a sentence, thus it should be given the same attention as the mean dependency distance (Chen and Gerdes 2020). What's more, power-law distribution, which was proposed by Zipf (1936) to describe word frequencies and its ranks, has been repeatedly observed in various linguistic units such as morpheme length, word length, sentence length, and so on (Pustet and Altmann 2005; Pande and Dhami 2012; Narisong et al. 2014). Distribution patterns have also been applied to model individual DDs in many languages and have been found to follow certain rules. Previous studies have revealed that the probability distribution of DDs fits Zipf-like laws well, especially the Zipf-Alekseev distribution (Jiang and Liu, 2015; Ouyang and Jiang 2017). Researchers also found that the Zipf-Alekseev function can well capture the distribution of many linguistic units of physical length, and that the parameters in the Zipf-Alekseev function reflect the peculiarities of human languages. For example, a language of an older stage has a larger parameter *a* than that of a younger one (Popescu et al. 2014); different genres demonstrate different parameters modeling dependency distance distribution (Wang and Liu, 2017); and the parameters are to some degree indicative of the proficiency of second language learners (Ouyang and Jiang 2017).

Liu (2008) and Futrell et al. (2015) investigated the distributions of DDs in different languages and discovered that Zipf-like law indicates a universal trend toward minimizing the DD in human languages, which is the phenomenon of dependency distance minimization (DDM). This finding connects DD to the short-term memory and "the least effort principle" (Zipf 1949) of human beings. To be more specific, long DDs are more difficult to process because they require more memory storage. However, since human short-term memory is limited, short DDs are preferable according to the least effort principle. The DDM has also been evidenced diachronically (Liu et al. 2022).

Although the "third code" belongs to natural language, translation activity is distinct from other language activities. In light of this, it is necessary to further verify whether the Zipf-Alekseev function holds true for the distribution of DD in translational language and whether there are differences between translational language and native language in regards to parameters fitting the Zipf-Alekseev function. This study therefore hypothesizes that, like other natural languages, translational language follows the Zipf-Alekseev function, but as a constrained language, its specific features differ from them.

Another gap is that previous research on DD analysis of translational language has rarely examined an examination of individual dependency types. Jiang and Liu (2018) have noticed that dependency types can reveal syntactic information in more detail, be it in cross-language or cross-genre comparisons. According to a previous study, the distribution of dependency type *nsubj*, which is a typical dependency type, may be a useful metric for distinguishing specific genres (Wang and Yan 2018). Thus, the analysis of individual dependency types can help us further delve into the fine-grained features of translational language.

There are other inadequacies in terms of corpus construction in previous research. Although Fan and Jiang (2019) compared translational language with its source text and comparable native text, they did not take translation directionality into consideration as they gathered language materials merely in one translation direction, namely the translation from Chinese into English. Nevertheless, some language features of translational language captured in one particular translation direction are not traceable in the opposite direction (House 2008). Therefore, to better examine the MDD and the distributions in translational language, it is necessary to observe the translation materials in both directions and compare the features of translational language with its source text and comparable native text.

Given the above gaps, this study makes intra- and inter-lingual comparisons in both translation directions. In this study, we selected a language pair from distinct language families, i.e., the Chinese-English language pair. We collected language materials from both translation directions, i.e., translations from Chinese into English and those from English into Chinese. Then, we built a corpus comprising parallel and comparable materials. After that, MDDs were computed and individual DDs were examined to capture distribution patterns of translated texts in both translation directions, and of their corresponding source texts and

comparable native texts. Finally, fine-grained features reflected by the most frequently used dependency type *nsubj* were inspected in terms of its distribution and parameters. The main research questions addressed in the study are as follows:

1. What are the differences in MDD among translated texts, corresponding source texts and comparable native texts in target language?
2. Does the probability distribution of DDs in the translated texts fit the Zipf-Alekseev distribution well? Do the fitting parameters of the translated texts differ from those of the source texts and native texts in both translation directions?
3. Does the DD distribution of typical dependency type *nsubj* follows the Zipf-Alekseev distribution? Are the fitting parameters of the translated texts different from those of the source texts and native texts in both translation directions?

The rest of this paper is organized as follows. Section 2 describes the language materials and quantitative methods employed. Section 3 presents the results and discussion, after which a brief conclusion is drawn.

Materials and methods

Bidirectional parallel and comparable corpus. Previous studies on translational language mainly focused on Indo-European language pairs, in which the two languages involved are from the same language family. Only a few studies dealt with distinct family pairs (Xiao 2010; Wang et al. 2023). Since features obtained from distinct language pairs might be more convincing for generalization, the language pair selected in current study is Chinese-English. These two languages belong to the Sino-Tibetan and Indo-European language families, respectively.

To comprehensively explore the distinctive features of translational language, we built a corpus comprising parallel and comparable materials in both translation directions. Firstly, language features of translational language captured in one particular translation direction need evidence in the opposite direction. Secondly, both source text and comparable native text should be taken as references to examine S-type and T-type features of translational language. S-type features are depicted by comparing the source text with its corresponding translated text, which could be achieved based on the parallel language materials, while T-type features are captured by comparing the translated text with its comparable native text in target language, which is generally achieved through comparable language materials (Chesterman 2010).

To achieve the research objective, we collected language materials from the original Chinese Report on the Work of the Government delivered at the annual China National People's Congress by premiers of China's State Council and its English translation, as well as American presidential State of the Union and its Chinese translation. To ensure the comparability of the materials, the two kinds of native texts are of similar formality and genre. What is more, the translation materials are reliable as they were translated by professional translators and proofread by native speakers. For clarity, these four types of text will be referred to as NATIVE-ch, TRANS-en, NATIVE-en and TRANS-ch.

The language materials in the self-built bidirectional parallel and comparable corpus span two decades of the 21st century from 2000 to 2018, with a total of approximately 840,000 tokens. The materials utilized to test the probability distribution of DD are from the same 4-year time consisting of 67,620 English word tokens and 117,396 Chinese characters, respectively.

Data analysis. According to the syntactic analysis approach of dependency grammar, the sentence is analyzed syntactically in terms of the dependency relations between each pair of words (Tesnière 1959; Nivre 2006; Hudson 2010). A dependency relation has three core properties: it is a binary relation between two linguistic units; it is usually asymmetrical with one of the two units acting as governor and the other as dependent; it is labeled and the type of a dependency relation is usually indicated using a label on top of the arc linking the two units (Liu 2009a).

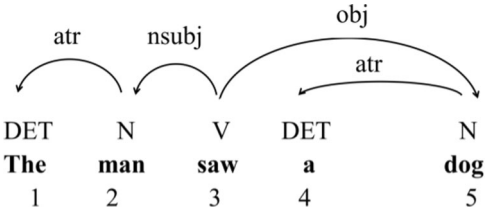

Fig. 1 Dependency tree of the sample sentence "*The man saw a dog*".

A directed dependency tree can be constructed based on the above properties, illustrating the syntactic structure of a sentence. The dependency trees of the sentence *The man saw a dog* and it's Chinese translation "那个男人看到一只狗" are shown in Figs. 1, 2. As shown in the figures, it is apparent the two sentences share the same structure.

Figures 1, 2 demonstrate the dependency relationships between governors and dependents in the sample sentences. Between each pair of words, a labeled arc with an arrow above the words points from the governor to the dependent. The labels above the arcs indicate dependency types, while the labels above the words are part of speech. The number below the words indicates the linear position or the order of the word in the sentence.

The linear distance between the governor and the dependent is defined as "dependency distance". The concept was first introduced by Heringer et al. (1980) while the term "dependency distance" was first used by Hudson (1995) and defined as "the distance between words and their parents, measured in terms of intervening words". Later, Liu (2009a) proposed an easier method for calculating the DD of sentences and texts. "Formally, let $W_1 \ldots W_i \ldots W_n$ be a word string. For any dependency relation between two words W_a and W_b, if W_a is the governor and W_b is its dependent, then the DD between them can be measured as the difference a–b; by this measure, adjacent words have a DD of 1. When a is greater than b, the DD is a positive number, which means that the governor is after the dependent; when a is smaller than b, the DD is a negative number and the governor precedes the dependent." However, when measuring the mean dependency distance, we need the absolute value of DD.

The MDD of a sentence can be obtained with the formula:

$$\mathrm{MDD\,(sentence)} = 1/(n-1)\sum_{i=1}^{n-1}\left|DD_i\right| \quad (1)$$

where n is the sentence length and DD_i represents the DD of the i-th syntactic relation in the sentence.

In a sentence, there is only one word that doesn't have a governor. That word is the root verb. The DD of this root verb is considered as zero. Thus, we can obtain several $|DD_i|$ for the sample sentence *The man saw a dog* as follows: $|DD_i| = 1\ 1\ 1\ 2$, which are obtained by subtracting the order number of the word from the order number of its governor. Then, by Formula (1), the MDD of the sample sentence is $5/4 = 1.25$.

Formula (2) can be used to calculate the MDD of a treebank:

$$\mathrm{MDD\,(treebank)} = 1/(n-s)\sum_{i=1}^{n-s}\left|DD_i\right| \quad (2)$$

where n is the number of words in the treebank and s is the number of sentences in the treebank. DD_i describes the DD of the i-th syntactic relation of the treebank.

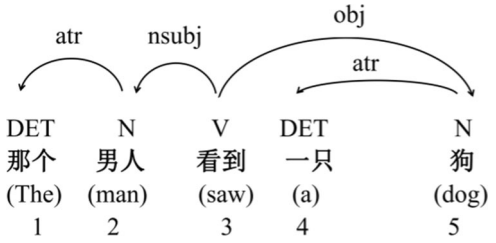

Fig. 2 Dependency tree of the sample sentence "那个男人看到一只狗".

After constructing the corpus, we segmented the Chinese texts into words using ICTCLAS (Zhang et al. 2003), a tool for segmenting Chinese text strings into word tokens. Then, we transformed the corpus into dependency-annotated treebanks. The texts were parsed with the Stanford Parser (3.6.0), a natural language processing program developed by Stanford University. The parsed results were manually checked then to make sure that the texts were parsed following the same annotation scheme. Finally, DDs and MDDs were computed.

Based on the distribution models in previous quantitative linguistic studies of distribution of DDs (Jiang and Liu 2015), it is hypothesized that the distribution of DDs in translational language obey the Zipf-Alekseev model (Hřebíček 1996). Two assumptions were adopted by Hřebíček:

(1) the logarithm of the ratio of the probabilities P_1 and P_x is proportional to the logarithm of the class size, i.e.

$$\ln\left(\frac{P_1}{P_x}\right) \propto \ln x \tag{3}$$

(2) the proportionality function is given by the logarithm of Menzerath's law (Hierarchy), i.e.

$$\ln\left(\frac{P_1}{P_x}\right) = \ln\left(Axe^b\right)\ln x \tag{4}$$

yielding the solution

$$P_x = P_1 x^{-(a+b\ln x)},\ x = 1, 2, 3, \ldots \tag{5}$$

If Eq. (3) is considered a probability distribution, then P_1 is the norming constant, otherwise it is estimated as the size of the first class, $x = 1$. Very often, diversification distributions display a diverging frequency in the first class while the rest of the distributions behave regularly. In these cases, one usually ascribes the first class a special value α, modifying Eq. (3) as

$$P_x = \begin{cases} \alpha, x = 1 \\ \frac{(1-a)x^{(a+n\ln x)}}{T}, x = 2,\ 3,\ \ldots,(n) \end{cases} \tag{6}$$

where

$$T = \sum_{j=2}^{n} j^{-(a+b\ln j)},\ a,\ b\varepsilon\Re, 0<\alpha<1$$

Distributions (5) and (6) are called Zipf-Alekseev distributions. If n is finite, Eq. (6) is called a Right truncated modified Zipf-Alekseev distribution. In our study, we used the Altmann-Fitter (Altmann-Fitter 2013) to fit the model to the data under study and obtain the goodness of fit and parameters.

Results and discussion

Mean dependency distance. MDD has been quantified in various languages as a global indicator of complexity and cognitive cost in the framework of dependency grammar. For instance, the MDD of Chinese is found to be relatively longer than that of any other languages such as English, Japanese and Italian (Liu 2008). To examine the global distinctive features of translational language, the MDDs of translated texts in both translation directions and of their source texts, as well as of their comparable native texts were calculated and summarized in Table 1.

Table 1 The MDD of four text types.

Text Type	MDD
NATIVE-ch	4.567
TRANS-ch	4.201
TRANS-en	3.713
NATIVE-en	3.017

As can be seen in the table, the MDD of translated Chinese texts (TRANS-ch) is longer than that of their English source texts (NATIVE-en), but shorter than that of their comparable native Chinese texts (NATIVE-ch). A one-way ANOVA and Games-Howell's post hoc tests confirmed the significance of the difference in MDD among the three types of texts, $F_{(2,54)} = 175.184$, $p < 0.001$. At the same time, the MDD of translated English texts (TRANS-en) is shorter than that of their Chinese source texts (NATIVE-ch), but longer than that of their comparable native English texts (NATIVE-en), $F_{(2,54)} = 130.044$, $p < 0.001$. From the data, it is evident that translated texts in both translation directions yield a compromising MDD in between their source texts and comparable native texts.

Interestingly, the in-between MDD of translated texts is consistent with the findings by Fan and Jiang (2019), in which evidence for the in-between MDD of translated texts presented itself in Chinese-English translation. In that study, the feature was attributed to the negotiation between the source language and the target language during the translating process. During the process, the source text activates the source language processing system in the brain, which in turn affects the target text production. Both the source language and the target language systems are simultaneously activated in the brain (Mauranen 2004). Thus, we can infer that the higher MDD in English translations than native English might be attributed to the bigger cognitive effort in dealing with the Chinese source texts.

Probability distribution of DDs. As mentioned in the first section of this paper, the features of translational language need to be explored in individual dependency distances, which may provide more details about the fluctuation than MDD. Constrained by human working memory capacity, natural languages tend to minimize DD, which causes the distribution of DDs to follow certain Zipf-like laws. To answer the second research question, the individual dependency distances of the four types of texts were extracted. The data of these four types were then fitted to the Right truncated modified Zipf-Alekseev by Altmann-Fitter, and the goodness-of-fit and parameters are listed in Table 2. The obtained distributions of the DDs of four types of texts and their log-log format are presented in Figs. 3–6.

In Figs. 3–6, the four curves of the original data illustrate the relationship between the DDs of the four types and their frequencies. All the four curves show a sharp decline followed by a long flat tail, which demonstrates power-law distributions. As for the fitting results, the determination coefficient R^2 is generally regarded as a criterion for evaluating the goodness-of-fit (Liu 2009b). The formula of R^2 is defined as

$$R^2 = \sum_{i=1}^{n} \frac{\left(f_i - NP_i\right)^2}{NP_i} \tag{7}$$

where f_i is the observed frequency of the value i, P_i the expected probability of the value i, n the number of different data values, and N the sample size.

Table 2 The parameters of fitting result.

Text	*a*	*b*	*α*	X^2	$P(X^2)$	*DF*	*n*	*C*	R^2
NATIVE-ch	0.4745	0.3900	0.4385	564.6408	0	81	116	0.0159	0.9954
TRANS-en	0.1678	0.4924	0.3639	2498.664	0	76	97	0.0408	0.9824
NATIVE-en	0.0214	0.6332	0.3904	761.3573	0	42	67	0.0307	0.9894
TRANS-ch	0.4836	0.4061	0.4195	119.6617	0	60	73	0.0056	0.9983

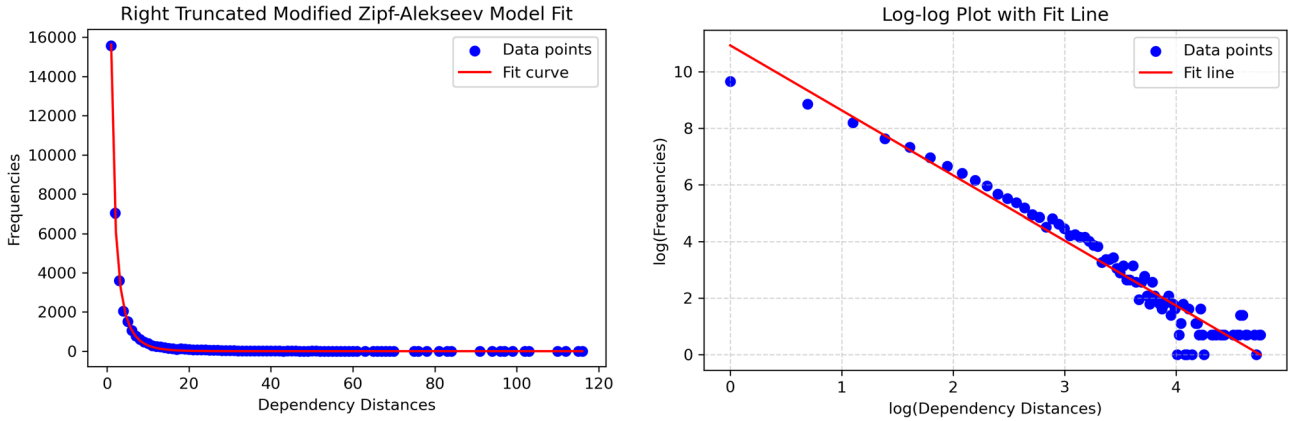

Fig. 3 DD Distribution of NATIVE-ch.

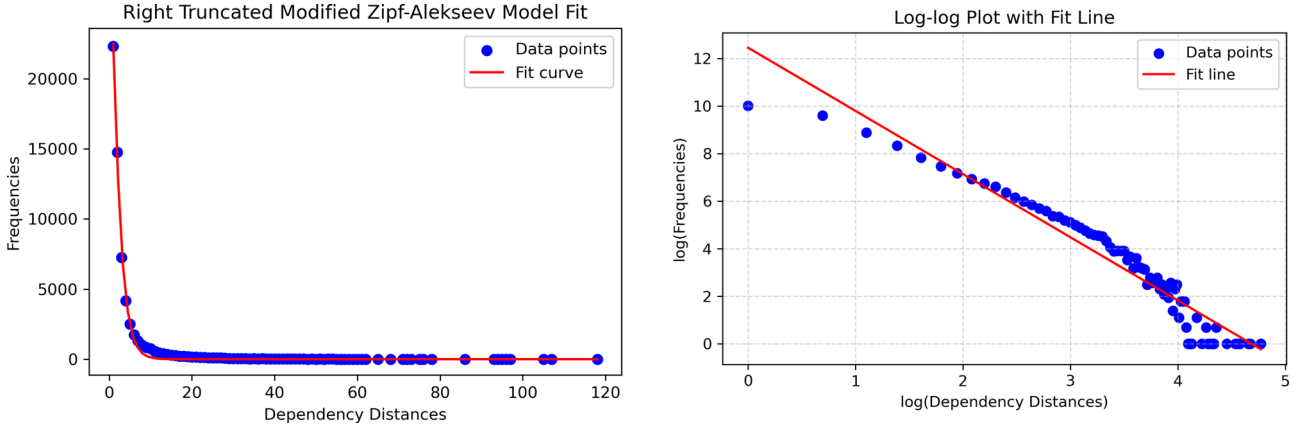

Fig. 4 DD Distribution of TRANS-en.

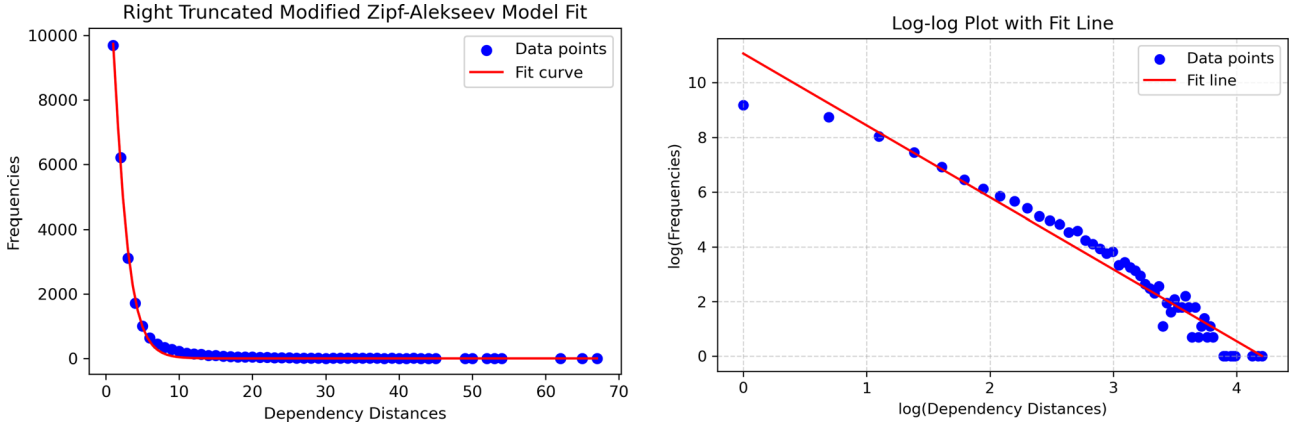

Fig. 5 DD Distribution of NATIVE-en.

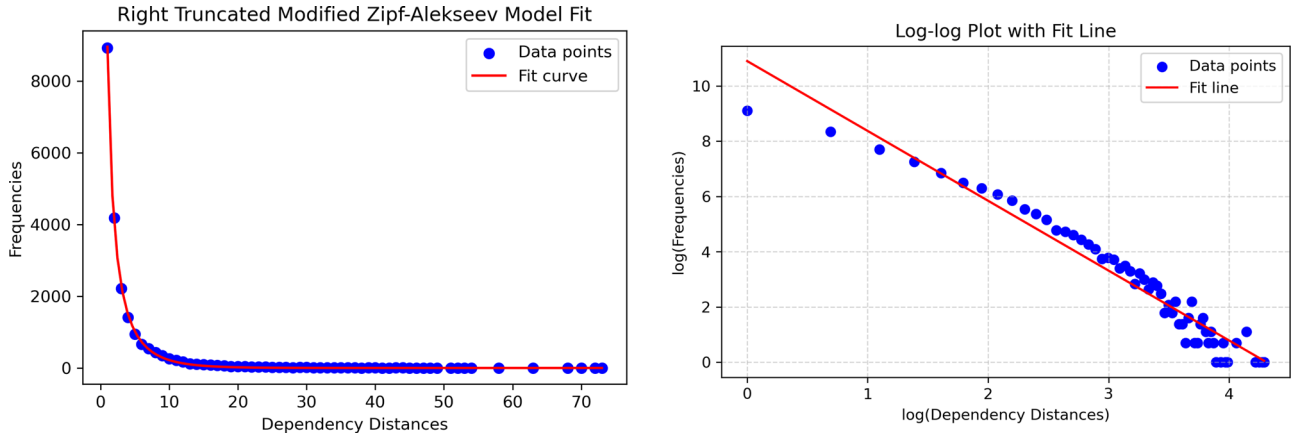

Fig. 6 DD Distribution of TRANS-ch.

Table 2 demonstrates that R^2 values of all the four types are above 0.98, indicating satisfactory fitting results of the observed data to the Right truncated modified Zipf-Alekseev distribution, a kind of power-law distribution. *C* stands for the coefficient of variation, and in general, the closer the *C* value is to zero, the better the result. Table 2 shows that the *C* values for these four curves are relatively small, all below 0.05. Therefore, although the MDDs of the four types of texts differ from one another, the distributions of the DDs all conform to a same power-law distribution. The results suggest that both native language or translational language fit the Zipf-Alekseev distribution well, providing a positive answer to the first part of the second research question.

Remarkably, the most frequent DD in all the four curves is the adjacent dependency distance, i.e., the shortest dependency distance, DD of 1. Firstly, the result validates the DDM phenomenon, suggesting that there is indeed a preference for short DDs in naturally produced languages due to cognitive or short-term memory constraints. Secondly, the result demonstrates that the DDM also works in producing translational language, suggesting that cognitive constraints act not only on native language producers but also on translators. Another notable observation is that there are DD values exceeding 100 in NATIVE-ch and TRANS-en. This phenomenon can be attributed to the stylistic features of the Chinese source texts. Government work reports often feature lengthy sentences with parallel constructions, such as "In response to…, we deepen…, accelerate…, reduce…, and emphasize…, …", contributing to extended sentence length and consequently longer DD between words. Moreover, it results in longer sentence length and DD in the translated English as well.

The distribution of DDs in both native and translational language follows a power-law distribution, indicating that it's possibly a feature shared by all human language. This is probably determined by human cognitive load capacity and "the least effort principle". That is to say, during the process of producing the translated texts, translators prefer short DDs to longer ones so as to decrease the cognitive load both for themselves and for the readers. It is worth noting that although translational language has features in many aspects different from native language, it shares similar regularities with native language, which may be regarded as language universals. Nevertheless, the quest for translation universals is also meaningful in that the cognitive process involved in the production of translational texts is different from that in the production of native language.

The fitting parameters of the Zipf-Alekseev distribution. Popescu et al. (2014) proposed that the parameters in the Zipf-Alekseev distribution might reflect the peculiarities of human languages and that the parameters themselves are part of a dynamic system displaying self-regulation. In order to investigate whether translational language in this adaptive dynamic system presents features different from its source texts and the corresponding native language with regards to fitting parameters, the current study conducted a comparative analysis of the fitting parameters across the four types of texts.

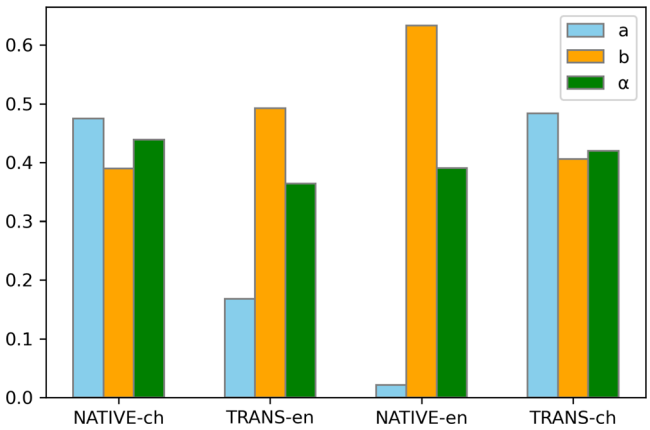

Fig. 7 Bar graph of the parameters *a*, *b* and *α*.

Table 2 shows the parameters of the four types of text fitting the Right truncated modified Zipf-Alekseev distribution. To visualize the comparison of the three crucial parameters *a*, *b* and *α*, we present them in a bar graph in Fig. 7. The Figure demonstrates that parameter *α* remains basically unchanged, fluctuating slightly around 0.4, which is consistent with the findings by Ouyang and Jiang (2017). Therefore, parameter *α* was not taken into consideration while investigating the distinctive features of translational language in this study. Next, parameters *a* and *b* will be analyzed and interpreted.

In the case of native texts, there are great differences between Chinese and English in parameters *a* and *b*. To be specific, the parameter *a* of the native Chinese texts is considerably larger than that of the native English texts, while the parameter *b* of the native Chinese texts is much smaller than that of the native English texts. According to Table 1, the MDD of native Chinese is longer than that of native English, mirroring the higher syntactic complexity of the Chinese texts than the English texts, evidence of which can also be found in Liu's research (Liu 2008). Returning to Fig. 7, we can infer that parameter *a* increases as syntactic complexity increases and parameter *b* decreases as syntactic

complexity increases, which matches the findings by Ouyang and Jiang (2017).

In the case of translated texts, we will interpret the result from two translation directions respectively. In Chinese-English translation, both parameters *a* and *b* in the translated texts show a compromise between the Chinese source texts and the English native texts. The two parameters, as components of the dynamic system, are subject to the gravitation pull from both the source and target languages, leading to the emergence of a "compromise feature", which is consistent with the results in MDD in the above section. Nevertheless, in English-Chinese translation, the values of the parameters *a* and *b* for translated Chinese and native Chinese texts are remarkably close to each other, indicating that Chinese materials are not sensitive to the two parameters.

To sum up, in terms of the fitting parameter *α*, the translated texts do not differ from the native texts. However, the "in-between feature" is verified in parameters *a* and *b* in the Chinese-English translation materials, while the English-Chinese translation materials are not sensitive to the two parameters. The "in-between feature" reveals that the translated English texts are characterized by a compromise between the Chinese source texts and the native English texts.

Probability distribution of the DDs of dependency type nsubj. As mentioned above, there have been relatively few studies on the distribution of the DDs of specific dependency types. To further explore the fine-grained features of translational language, this study investigates a high-frequency dependency type *nsubj* in terms of its distribution of DDs and fitting parameters. According to a previous study (Wang and Yan 2018), the dependency type *nsubj*, viz., nominal subject, representing the subject-predicate relationship in a sentence, is one of the most crucial dependency types typical of the vast majority of languages. To figure out whether the distribution of the DDs of dependency type *nsubj* follows the Zipf-Alekseev distribution and whether the parameters can reveal the difference between translational language and native language, we fitted the distribution of the DDs of dependency type *nsubj*.

The distributions of the DDs of dependency type *nsubj* in the four types of texts and their log-log format are presented in Figs. 8–11. The data of the four types of texts were then fitted to the Right truncated modified Zipf-Alekseev distribution. The goodness-of-fit and parameters are shown in Table 3.

Figures 8–11 show that all the four distribution curves of the dependency type *nsubj* present a sharp decline followed by a long flat tail, demonstrating a power-law distribution. In response to the third research question, Figs. 8–11 indicate that the distribution of DDs of the typical dependency type *nsubj* follows the Zipf-Alekseev distribution in both native and translational languages. However, a closer inspection of the figures reveals that the DD distribution curve of the translated English texts (TRANS-en) is slightly different from the other three curves. The curve of TRANS-en obviously deviates from the other curves, with the most frequent DD that appears in this type of texts being

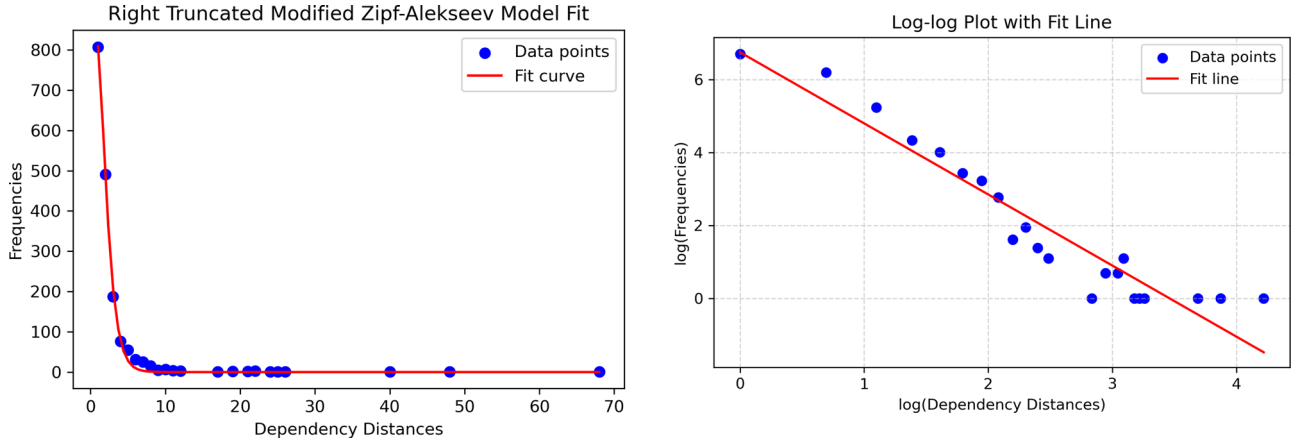

Fig. 8 DD Distribution of dependency type *nsubj* in NATIVE-ch.

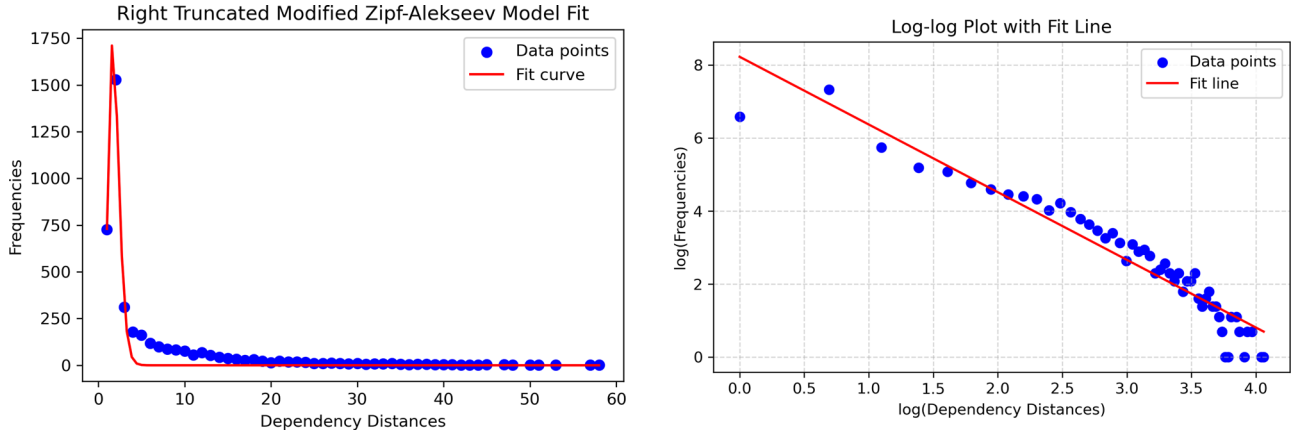

Fig. 9 DD Distribution of dependency type *nsubj* in TRANS-en.

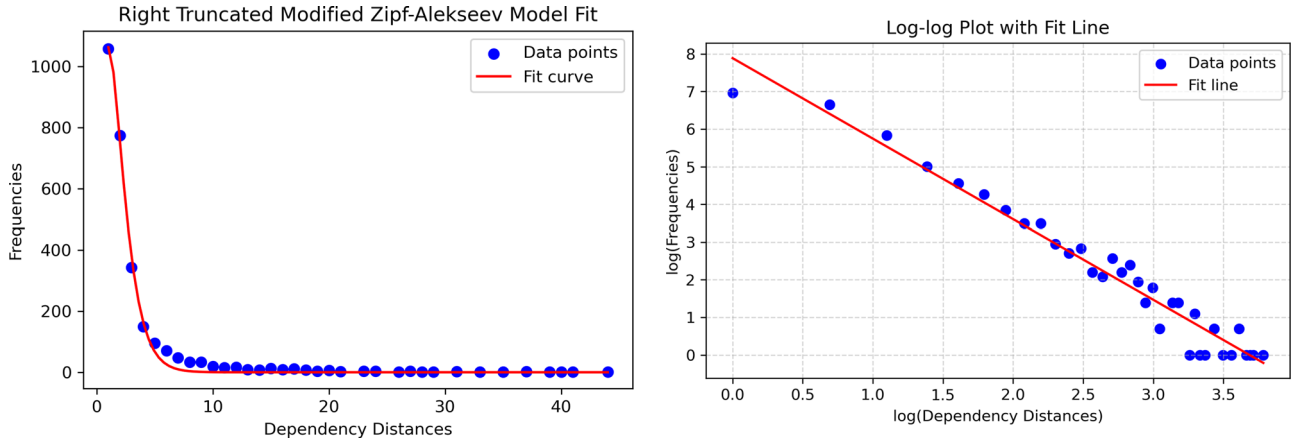

Fig. 10 DD Distribution of dependency type *nsubj* in NATIVE-en.

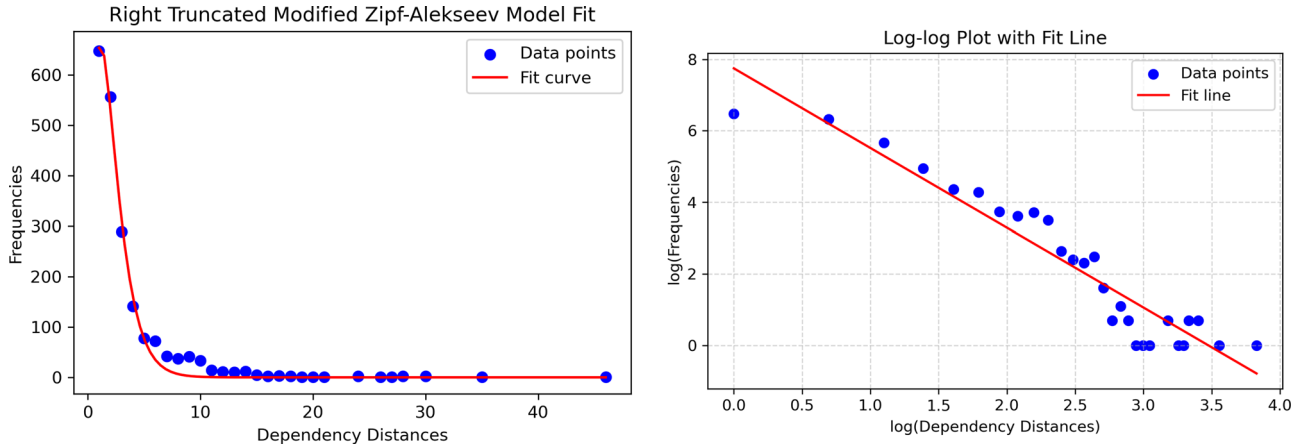

Fig. 11 DD Distribution of dependency type *nsubj* in TRANS-ch.

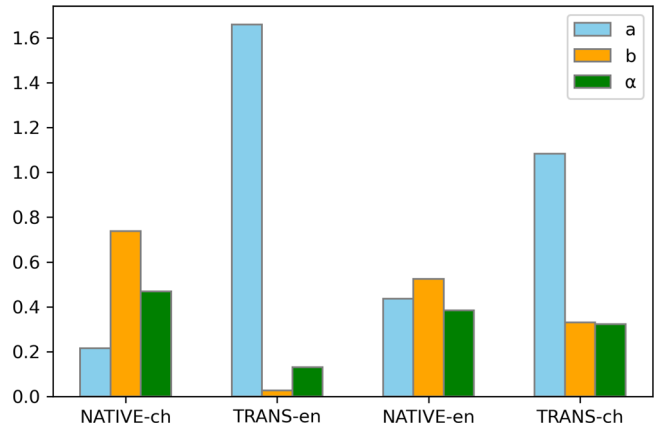

Fig. 12 Bar graph of the parameters *a*, *b* and *α* in fitting result of dependency type *nsubj*.

the DD of 2, rather than DD of 1, i.e., the adjacent dependency distance. Furthermore, Table 3 illustrates that the determination coefficient R^2 of the other three types of texts are all above 0.98, while the R^2 of TRANS-en is 0.9769. Although all the curves fit the Right truncated modified Zipf-Alekseev distribution well, TRANS-en shows a lower degree of goodness-of-fit than the other three types.

To find a possible explanation for the divergence mentioned above, we delved deeper into the corpus. In the corpus, it is evident that tenses in Chinese are not expressed as explicitly as in English, which results in frequent omissions of tense markers in Chinese. Meanwhile, Chinese syntax is relatively flexible, so that the subject is often omitted as well. For instance, "加大补短板力度(jia da bu duan ban li du) (Literal translation: step up the effort to strengthen areas of weakness)" and "加快新旧发展动能接续转换(jia kuai xin jiu fa zhan dong neng jie xu zhuan huan) (Literal translation: speed up the replacement of old growth drivers)" are two sentences in NATIVE-ch with "verb+object" structure where the subject is omitted. However, they were translated in English in the parallel corpus as "We have stepped up efforts to strengthen areas of weakness" and "We have sped up the replacement of old growth drivers". It can be seen in the translated English text that the tense markers were explicitly manifested and the omitted subjects were added. This can be attributed to the fact that tense markers in English cannot generally be omitted and that the complete syntactic structure of SVO is strictly required upon most occasions. Therefore, there are plenty of expressions in TRANS-en such as "we will do…" and "we have done…", which correspondingly leads directly to a drastic increase in the frequency of DD of 2 in the case of dependency type *nsubj*.

Table 3 lists the parameters of the Right truncated modified Zipf-Alekseev distribution fitted by the distribution of DD of dependency type *nsubj* in the four types of texts. A bar graph in

Table 3 The parameters of fitting result of dependency type *nsubj*.

Text	*a*	*b*	*α*	X^2	$P(X^2)$	*DF*	*n*	*C*	R^2
NATIVE-ch	0.2145	0.7375	0.4689	88.8256	0	14	68	0.0516	0.9862
TRANS-en	1.6588	0.0267	0.1298	223.6833	0	53	58	0.0399	0.9769
NATIVE-en	0.4366	0.5245	0.3852	136.0573	0	25	44	0.0495	0.9823
TRANS-ch	1.0819	0.3303	0.3225	45.9985	0	28	46	0.0229	0.9977

Fig. 12 presents a comparison of the parameters *a*, *b* and *α*. As can be seen in the figure, parameter *α* agrees with the result in Section 3.3 that it doesn't vary notably with the type of text. Nevertheless, the other two parameters, especially parameter *a*, exhibit great variation with translation status. To be specific, parameter *a* is notably larger in both translated English and translated Chinese than in native English and native Chinese, while parameter *b* shows smaller values in both translated texts, especially in translated English, than in native English and native Chinese. To conclude, the bar graph reveals that parameter *a* is larger in the translated texts of both translation directions than their source texts and comparable native texts in the target language, and that parameter *b* is smaller in the translated texts of both translation directions than their source texts and native texts. According to a previous study (Ouyang and Jiang 2017), parameter *a* increases as syntactic complexity increases while parameter *b* decreases as the complexity increases. Therefore, for the current study, translational language is syntactically more complex than source and native languages in terms of the dependency type *nsubj*, which is not dependent upon a specific direction of translation, no matter whether it is C-E or E-C translation. This conclusion suggests that dependency type *nsubj* is probably crucially involved in shaping the structure of translational language.

Conclusion

Translational language has been referred as the "third code" due to its distinctive features. To further explore its features, the present study adopts the approach of quantitative linguistics to investigate the MDD, the probability distribution of individual DDs and that of a typical dependency type in translational language. The MDD and distributions were tested in a bidirectional parallel and comparable corpus in which both intra- and inter-lingual comparisons were conducted.

The results are in the first place consistent with the prior studies (Fan and Jiang 2019, 2020), showing that in both translation directions, translated texts yield a compromising MDD in between their source texts and comparable native texts. Moreover, the results demonstrate that both native language and translational language fit the Zipf-Alekseev distribution well. The "in-between feature" is further verified by parameters *a* and *b* in the Chinese-English translation materials, while the English-Chinese translation materials are not sensitive enough to the two parameters.

In addition, the distribution of DDs of the typical dependency type *nsubj* follows the Zipf-Alekseev distribution in both native language and translational language. Nevertheless, parameters *a* and *b* in the distribution vary drastically with the change in translation status. Specifically, in both translation directions, parameter *a* is larger in translated texts than their source texts and native texts in the target language, while parameter *b* displays an opposite trend.

The results reveal that, on the one hand, the distributions confirm the existence of the DDM phenomenon, suggesting that there is indeed a preference for short DDs in naturally produced language due to human cognitive constraints and short-term memory limitations. Furthermore, the DDM also works in the production of translational language, further evidencing that cognitive constraints act not only on native language producers but also on translators. On the other hand, the parameters indicate that there is a "in-between feature", which frequently occurs in translational language, independent of translation directions.

By answering the three research questions, our study confirms that translational language does have some features that are clearly different from native language. The study also illustrates the potential of syntactic quantitative methods for translation studies. Nevertheless, it is worth noting that in a prior study (Liu, 2008), the MDD is 3.662 for Chinese and 2.543 for English, while in our study, the MDDs for Chinese and English, no matter in translated or native texts, are higher than those in Liu's study. This may probably result from the fact that the genre of the materials in this study is political news, containing formally-published government documents. Therefore, future studies are warranted to verify the results in materials of other genres.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 10 July 2023; Accepted: 20 November 2023;
Published online: 06 December 2023

References

Altmann-Fitter (2013) Altmann-Fitter user guide. The third version. Retrieved August 29, 2016, from http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-User-Guide.pdf
Baker M (1996) Corpus-based translation studies: the challenges that lie ahead. In: Somers H (ed) Terminology, LSP and Translation: studies in language engineering: in honour of Juan C. Sager. John Benjamins, Amsterdam and Philadelphia, p 175–186
Chen XY, Gerdes K (2020) Dependency distances and their frequencies in Indo-European Language. J Quant Linguist 1:1–19
Chesterman A (2010) Why study translation universals? Acta Translatologica Helsingiensia 1:38–48
Duff A (1981) The Third Language: recurrent problems of translation into English: it ain't what you do, it's the way you do it. Pergamon Press, Oxford
Fan L, Jiang Y (2019) Can dependency distance and direction be used to differentiate translational language from native language? Lingua 224:51–59
Fan L, Jiang Y (2020) A syntactic dependency network approach to the study of translational language. Digital Scholarship Humanit 3:595–606
Ferrer-i-Cancho R (2015) The placement of the head that minimizes online memory: a complex systems approach. Lang Dyn Change 1:114–137
Frawley W (1984) Prolegomenon to a theory of translation. In: Frawley W (ed) Translation: literary, linguistic, and philosophical perspectives. Associated University Presses, London, p 159–175
Futrell R, Mahowald K, Gibson E (2015) Large-scale evidence of dependency length minimization in 37 languages. Proc Natl Acad Sci 33:10336–10341
Heringer HJ, Strecker B, Wimmer R (1980) Syntax Fragen-Lösungen Alternativen. Wilhelm Fink Verlag, München
House J (2008) Beyond intervention: universals in translation? trans-kom 1:6–19
Hřebíček L (1996) Word associations and text. Wissenschaftlicher Verlag, Trier

Hudson R (1995) Measuring syntactic difficulty. Unpublished paper. Retrieved October 8, 2017 from http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf
Hudson R (2010) An Introduction to Word Grammar. Cambridge University Press, Cambridge
Jiang JY, Liu HT (2015) The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. Lang Sci 50:93–104
Jiang JY, Liu HT (2018) Quantitative Analysis of Dependency Structures. De Gruyter, Berlin and Boston
Kruger H, Rooy BV (2016) Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. Engl World-Wide 1:26–57
Laviosa S, Liu KL (2021) The Pervasiveness of Corpora in Translation Studies. Transl Quart 101:5–20
Liu HT (2008) Dependency distance as a metric of language comprehension difficulty. J Cognit Sci 2:159–191
Liu HT (2009a) Dependency Grammar: from theory to practice. Science Press, Beijing
Liu HT (2009b) Probability distribution of dependencies based on a Chinese dependency Treebank. J Quant Linguistics 3:256–273
Liu HT, Xu CS, Liang JY (2017) Dependency distance: A new perspective on syntactic patterns in natural languages. Phys Life Rev 21:171–193
Liu KL, Liu ZZ, Lei L (2022) Simplification in translated Chinese: An entropy-based approach. Lingua 275:103364
Liu XY, Zhu HR, Lei L (2022) Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types. Human Soc Sci Commun 1:420
Mauranen A (2004) Corpora, universals and interference. In: Mauranen A, Kujamäki P (eds) Translation Universals: Do They Exist? John Benjamins, Amsterdam and Philadelphia, p 65–82
Narisong, Jiang JY, Liu HT (2014) Word length distribution in Mongolian. J Quant Linguistics 2:123–152
Nivre J (2006) Inductive Dependency Parsing. Springer, Dordrecht
Ouyang JH, Jiang JY (2017) Can the probability distribution of dependency distance measure language proficiency of second language learners? J Quant Linguistics 4:1–19
Pande H, Dhami HS (2012) Model generation for word length frequencies in texts with the application of Zipf's order approach. J Quant Linguistics 4:249–261
Popescu II, Best KH, Altmann G (2014) Unified modeling of length in language (Studies in quantitative linguistics 16). RAM-Verlag, Lüdenscheid
Pustet R, Altmann G (2005) Morpheme length distribution in Lakota. J Quant Linguistics 1:53–63
Tesnière L (1959) Éléments de syntaxe structurale. Klincksieck, Paris
Trosborg A (1997) Translating hybrid political texts. In: Trosborg A (ed) Text Typology and Translation. John Benjamins, Amsterdam and Philadelphia, p 119–143
Wang YQ, Liu HT (2017) The effects of genre on dependency distance and dependency direction. Lang Sci 59:135–147
Wang YQ, Yan JW (2018) A quantitative analysis on a literary genre essay's syntactic features. In: Jiang JY, Liu HT (eds) Quantitative analysis of dependency structures. de Gruyter, Berlin and Boston, p 295–314
Wang ZL, Liu KL, Moratto R (2023) A corpus-based study of syntactic complexity of translated and non-translated chairman's statements. Transl Interpreting 15:135–151
Xiao R (2010) How different is translated Chinese from native Chinese? Int J Corpus Linguistics 1:5–35
Zhang HP et al. (2003) HHMM-based Chinese lexical analyzer ICTCLAS. Proceedings of the second SIGHAN workshop on Chinese language processing-Volume17. Association for Computational Linguistics
Zipf GK (1936) The psychobiology of language. Routledge, London
Zipf GK (1949) Human behavior and the principle of least effort: an introduction to human ecology. Addison-Wesley Press, Cambridge

Acknowledgements

This work is supported by the National Social Science Foundation of China (Grant No.22CYY024).

Author contributions

LF contributed to the design and implementation of the work. All authors were involved in the interpretation of data for the work. LF drafted the work and YJ revised and proofread it. All authors approved the final version.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed Consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-02427-x.

Correspondence and requests for materials should be addressed to Lu Fan.